Tensor Graphical Model: Non-convex Optimization and Statistical Inference

Xiang Lyu, Will Wei Sun, Zhaoran Wang, Han Liu, Jian Yang, Guang Cheng

Abstract—We consider the estimation and inference of graphical models that characterize the dependency structure of high-dimensional tensor-valued data. To facilitate the estimation of the precision matrix corresponding to each way of the tensor, we assume the data follow a tensor normal distribution whose covariance has a Kronecker product structure. A critical challenge in the estimation and inference of this model is the fact that its penalized maximum likelihood estimation involves minimizing a non-convex objective function. To address it, this paper makes two contributions: (i) In spite of the non-convexity of this estimation problem, we prove that an alternating minimization algorithm, which iteratively estimates each sparse precision matrix while fixing the others, attains an estimator with an optimal statistical rate of convergence. (ii) We propose a de-biased statistical inference procedure for testing hypotheses on the true support of the sparse precision matrices, and employ it for testing a growing number of hypothesis with false discovery rate (FDR) control. The asymptotic normality of our test statistic and the consistency of FDR control procedure are established. Our theoretical results are backed up by thorough numerical studies and our real applications on neuroimaging studies of Autism spectrum disorder and users' advertising click analysis bring new scientific findings and business insights. The proposed methods are encoded into a publicly available R package **Tlasso**.

Index Terms—asymptotic normality, hypothesis testing, optimality, rate of convergence.

1 Introduction

High-dimensional tensor-valued data are observed in many fields such as personalized recommendation systems and imaging research [1], [2], [3], [4], [5], [6], [7], [8], [9]. Traditional recommendation systems are mainly based on the user-item matrix, whose entry denotes each user's preference for a particular item. To incorporate additional information into the analysis, such as the temporal behavior of users, we need to consider tensor data, e.g., user-item-time tensor. For another example, functional magnetic resonance imaging (fMRI) data can be viewed as a three-way tensor since it contains brain measurements taken on different locations over time under various experimental conditions. Also, in the example of microarray study for aging [10], thousands of gene expression measurements are recorded on 16 tissue types on 40 mice with varying ages, which forms a four-way gene-tissue-mouse-age tensor.

In this paper, we study the estimation and inference of conditional independence structure within tensor data. For example, in the microarray study for aging we are interested in the dependency structure across different genes, tissues, ages and even mice. Assuming data are drawn from a tensor normal distribution, a straightforward way to estimate this structure is to vectorize the tensor and estimate the underlying Gaussian graphical model associated with the vector. Such an approach ignores the tensor structure and requires estimating a rather high dimensional precision matrix with an insufficient sample size. For instance, in the aforementioned fMRI application the sample size is one if we aim to estimate the dependency structure across different locations, time and experimental conditions. To address such a problem, a popular approach is to assume the covariance matrix of the tensor normal distribution is separable in the sense that it is the Kronecker product of small covariance matrices, each of which corresponds to one way of the tensor. Under this assumption, our goal is to estimate the precision matrix corresponding to each way of the tensor and recover its support. See §1.1 for a detailed survey of previous work.

The separable normal assumption imposes non-convexity on the penalized negative log-likelihood function. However, most existing literatures do not fix this gap between computational and statistical theory. As we will show in §1.1, previous work mainly focus on establishing the existence of a local optimum, rather than offering efficient algorithmic procedures that provably achieve the desired local optima. In contrast, we analyze an alternating minimization algorithm, named as Tlasso, that attains a consistent estimator after only one iteration. This algorithm iteratively minimizes the non-convex objective function with respect to each individual precision matrix while fixing the others.

The established theoretical guarantees of the Tlasso algorithm are as follows. Suppose that we have n observations from a K order tensor normal distribution. We denote by

- *X. Lyu and G. Cheng are with the Department of Statistics, Purdue University, West Lafayette, IN 47906, USA (e-mail: lyu17@purdue.edu, chengg@purdue.edu).*
- *W. Sun is with the Department of Management Science, University of Miami, FL 33146, USA (e-mail: wsun@bus.miami.edu).*
- *Z. Wang is with the Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208, USA (e-mail: zhaoranwang@gmail.com).*
- *H. Liu is with the Department of Electrical Engineering and Computer Science and the Department of Statistics, Northwestern University, Evanston, IL 60208 (e-mail: hanliu@northwestern.edu).*
- *J. Yang is with Yahoo Research, Sunnyvale, CA 94089 (e-mail: jianyang@oath.com).*

m_k, s_k, d_k $(k = 1, \dots, K)$ the dimension, sparsity, and max number of non-zero entries in each row of the k-th way precision matrix. Besides, we define $m = \prod_{k=1}^{K} m_k$. The k-th precision matrix estimator from the Tlasso algorithm achieves a $\sqrt{m_k(m_k + s_k)\log m_k/(nm)}$ convergence rate in Frobenius norm, which is minimax-optimal in the sense it is the optimal rate one can obtain even when the rest $K - 1$ true precision matrices are known [11]. Moreover, under an extra irrepresentability condition, we establish a $\sqrt{m_k \log m_k/(nm)}$ convergence rate in max norm, which is also optimal, and a $d_k\sqrt{m_k \log m_k/(nm)}$ convergence rate in spectral norm. These estimation consistency results, together with a sufficiently large signal strength condition, further imply the model selection consistency of edge recovery. Notably, these results demonstrate that, when $K \geq 3$, the Tlasso algorithm achieves above estimation consistency even if we only have access to one tensor sample, which is often the case in practice. This phenomenon was never observed in the previous work.

The dependency structure in tensor makes support recovery very challenging. To the best of our knowledge, no previous work has been established on tensor precision matrix inference. In contrast, we propose a multiple testing method. This method tests all the off-diagonal entries of precision matrix, built upon the estimator from the Tlasso algorithm. To further balance the performance of multiple tests, we develop a false discovery rate (FDR) control procedure. This procedure selects a sufficiently small critical value across all tests. In theory, the test statistic is shown to be asymptotic normal after standardization, and hence provides a valid way to construct confidence interval for the entries of interest. Meanwhile, FDR asymptotically converges to a pre-specific level. An interesting theoretical finding is that our testing method and FDR control are still valid even for any fixed sample size as long as dimensionality diverges. This phenomenon is mainly due to the utilization of tensor structure information corresponding to each mode.

In the end, we conduct extensive experiments to evaluate the numerical performance of the proposed estimation and testing procedures. Under the guidance of theory, we also propose a way to significantly accelerate the alternating minimization algorithm without sacrificing estimation accuracy. In the multiple testing method, we empirically justify the proposed FDR control procedure by comparing the results with the oracle inference results which assume the true precision matrices are known. Additionally, analyses of two real data, i.e., the Autism spectrum disorder neuroimaging data and advertisement click data from a major Internet company, are conducted, in which several interesting findings are revealed. For example, differential brain functional connectivities appear on postcentral gyrus, thalamus, and temporal lobe between autism patients and normal controls. Also, sports news and weather news are strongly dependent only on PC, while magazines are significantly interchained only on mobile.

1.1 Related Work and Contribution

A special case of our sparse tensor graphical model (when $K = 2$) is the sparse matrix graphical model, which is studied by [12], [13], [14], [15]. In particular, [12] and [13] only establish the existence of a local optima with desired statistical guarantees. Meanwhile, [14] considers an algorithm that is similar to ours. However, the statistical rates of convergence obtained by [13], [14] are much slower than ours when $K = 2$. See Remark 3.6 for a detailed comparison. For $K = 2$, our statistical rate of convergence in Frobenius norm recovers the result of [12]. In other words, our theory confirms that the desired local optimum studied by [12] not only exists, but is also attainable by an efficient algorithm. In addition, for matrix graphical models, [15] establishes the statistical rates of convergence in spectral and Frobenius norms for the estimator attained by a similar algorithm. Their results achieve estimation consistency in spectral norm with only one matrix observation. However, their rate is slower than ours with $K = 2$. See Remark 3.12 for detailed discussions. Furthermore, we allow K to increase and establish estimation consistency even in Frobenius norm for $n = 1$. Most importantly, all these results focus on matrix graphical model and can not handle the aforementioned motivating applications such as the gene-tissue-mouse-age tensor dataset.

In the context of sparse tensor graphical model with a general K, [16] show the existence of a local optimum with desired rates, but do not prove whether there exists an efficient algorithm that provably attains such a local optimum. In contrast, we prove that our alternating minimization algorithm achieves an estimator with desired statistical rates. To achieve it, we apply a novel theoretical framework to consider the population and sample optimizers separately, and then establish the one-step convergence for the population optimizer (Theorem 3.1) and the optimal rate of convergence for the sample optimizer (Theorem 3.4). A new concentration result (Lemma S.1) is developed for this purpose, which is also of independent interest. Moreover, we establish additional theoretical guarantees including the optimal rate of convergence in max norm, the estimation consistency in spectral norm, and the graph recovery consistency of the proposed sparse precision matrix estimator.

In addition to the literature on graphical models, our work is also related to another line of work about nonconvex optimization problems. See, e.g., [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29] among others. These existing results mostly focus on problems such as dictionary learning, phase retrieval and matrix decomposition. Hence, our statistical model and analysis are completely different from theirs.

Our work also connects with a recent line of work on Bayesian tensor factorization [30], [31], [32], [33], [34], [35], [36]. In particular, they model covariance structure along each mode of a single tensor as an intermediate step in their tensor factorization. These covariance structures are imposed on core tensor or factor matrices to serve as the priors. Our work is fundamentally different from these procedures as they focus on the accuracy of tensor factorization while we focus on the graphical model structure within tensor-variate data. In addition, their tensor factorization is applied on a single tensor while our procedure learns dependency structure of multiple high-dimensional tensor-valued data.

In the end, the tensor inference part of our work is related to the recent high dimensional inference work, [37],

[38] and [39]. The other two related work are [40] and [41]. To consider the statistical inference in the vector-variate high-dimensional Gaussian graphical model, [42] proposes the multiple testing procedure with FDR control, [43] extend the de-biased estimator to precision matrix estimation, and [44] consider a scaled-Lasso-based inference procedure. To extend the inference methods from the vector-variate Gaussian graphical model to the matrix-variate Gaussian graphical model, [45], [46] propose multiple testing methods with FDR control and establish their asymptotic properties. However, these existing inference work can not be directly applied to our tensor graphical model.

Notation: In this paper, scalar, vector and matrix are denoted by lowercase letter, boldface lowercase letter and boldface capital letter, respectively. For a matrix $\mathbf{A} = (\mathbf{A}_{i,j}) \in \mathbb{R}^{d\times d}$, we denote $\|\mathbf{A}\|_\infty, \|\mathbf{A}\|_2, \|\mathbf{A}\|_F$ as its max, spectral, and Frobenius norm, respectively. We define $\|\mathbf{A}\|_{1,\text{off}} := \sum_{i\neq j}|\mathbf{A}_{i,j}|$ as its off-diagonal ℓ_1 norm and $|||\mathbf{A}|||_\infty := \max_i \sum_j |\mathbf{A}_{i,j}|$ as the maximum absolute row sum. We denote $\text{vec}(\mathbf{A})$ as the vectorization of $\mathbf{A}$ which stacks the columns of the matrix $\mathbf{A}$. Let $\text{tr}(\mathbf{A})$ be the trace of $\mathbf{A}$. For an index set $\mathbb{S} = \{(i,j), i,j \in \{1,\ldots,d\}\}$, we define $[\mathbf{A}]_{\mathbb{S}}$ as the matrix whose entry indexed by $(i,j)\in\mathbb{S}$ is equal to $\mathbf{A}_{i,j}$, and zero otherwise. For two matrices $\mathbf{A}_1 \in \mathbb{R}^{m\times n}, \mathbf{A}_2 \in \mathbb{R}^{p\times q}$, we denote $\mathbf{A}_1 \otimes \mathbf{A}_2 \in \mathbb{R}^{mp\times nq}$ as the Kronecker product of $\mathbf{A}_1$ and $\mathbf{A}_2$. We denote $\mathbb{1}_d$ as the identity matrix with dimension $d\times d$. Throughout this paper, we use $C, C_1, C_2, \ldots$ to denote generic absolute constants, whose values may vary from line to line.

Organization: §2 introduces the main result of sparse tensor graphical model and its efficient implementation, followed by the theoretical study of the proposed estimator in §3. §4 contains all the statistical inference results including a novel test statistic for constructing confidence interval and a multiple testing procedure with FDR control. §5 demonstrates the superior performance of the proposed methods and performs extensive comparisons with existing methods in both parameter estimation and statistical inference. §6 illustrates analyses of two real data sets, i.e., the Autism spectrum disorder neuroimaging data and advertisement click data from a major Internet company, via the proposed testing method. §7 summarizes this article and points out a few interesting future work. Detailed technical proofs are available in supplementary material.

2 Tensor Graphical Model

This section introduces our sparse tensor graphical model and an alternating minimization algorithm for solving the associated nonconvex optimization problem.

2.1 Preliminary

We first introduce the preliminary background on tensors and adopt the notations used by [47]. Throughout this paper, higher order tensors are denoted by boldface Euler script letters, e.g. $\mathcal{T}$. We consider a K-way tensor $\mathcal{T} \in \mathbb{R}^{m_1\times m_2\times\cdots\times m_K}$. When $K = 1$ it reduces to a vector and when $K = 2$ it reduces to a matrix. The $(i_1,\ldots,i_K)$-th element of the tensor $\mathcal{T}$ is denoted as $\mathcal{T}_{i_1,\ldots,i_K}$. We denote the vectorization of $\mathcal{T}$ as $\text{vec}(\mathcal{T}) := (\mathcal{T}_{1,1,\ldots,1},\ldots,\mathcal{T}_{m_1,1,\ldots,1},\ldots,\mathcal{T}_{1,m_2,\ldots,m_K},\mathcal{T}_{m_1,m_2,\ldots,m_K})^\top \in \mathbb{R}^m$ with $m = \prod_k m_k$. In addition, we define the Frobenius norm of a tensor $\mathcal{T}$ as

$$\|\mathcal{T}\|_F := \sqrt{\sum_{i_1,\ldots,i_K} \mathcal{T}^2_{i_1,\ldots,i_K}}.$$

In tensors, a fiber refers to a higher order analogue of matrix row and column. A fiber is obtained by fixing all but one of the indices of the tensor, e.g., for a tensor $\mathcal{T}$, the mode-k fiber is given by $\mathcal{T}_{i_1,\ldots,i_{k-1},:,i_{k+1},\ldots,i_K}$. Matricization, also known as unfolding, is a process to transform a tensor into a matrix. We denote $\mathcal{T}_{(k)}$ as the mode-k matricization of a tensor $\mathcal{T}$. It arranges the mode-k fibers to be the columns of the resulting matrix. Another useful operation in tensor is the k-mode product. The k-mode (matrix) product of a tensor $\mathcal{T} \in \mathbb{R}^{m_1\times m_2\times\cdots\times m_K}$ with a matrix $\mathbf{A} \in \mathbb{R}^{J\times m_k}$ is denoted as $\mathcal{T}\times_k \mathbf{A}$ and is of the size $m_1\times\cdots\times m_{k-1}\times J\times m_{k+1}\times\cdots\times m_K$. Its entry is defined as

$$(\mathcal{T}\times_k \mathbf{A})_{i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K} := \sum_{i_k=1}^{m_k} \mathcal{T}_{i_1,\ldots,i_K}\mathbf{A}_{j,i_k}.$$

Furthermore, for a list of matrices $\{\mathbf{A}_1,\ldots,\mathbf{A}_K\}$ with $\mathbf{A}_k \in \mathbb{R}^{m_k\times m_k}$, we define

$$\mathcal{T}\times\{\mathbf{A}_1,\ldots,\mathbf{A}_K\} := \mathcal{T}\times_1\mathbf{A}_1\times_2\cdots\times_K\mathbf{A}_K.$$

2.2 Statistical Model

A tensor $\mathcal{T} \in \mathbb{R}^{m_1\times m_2\times\cdots\times m_K}$ follows the tensor normal distribution with zero mean and covariance matrices $\mathbf{\Sigma}_1,\ldots,\mathbf{\Sigma}_K$, denoted as $\mathcal{T} \sim \text{TN}(\mathbf{0};\mathbf{\Sigma}_1,\ldots,\mathbf{\Sigma}_K)$, if its probability density function is $p(\mathcal{T}|\mathbf{\Sigma}_1,\ldots,\mathbf{\Sigma}_K) =$

$$(2\pi)^{\frac{-m}{2}}\left\{\prod_{k=1}^K |\mathbf{\Sigma}_k|^{\frac{-m}{2m_k}}\right\}\exp\big(-\|\mathcal{T}\times\mathbf{\Sigma}^{\frac{-1}{2}}\|_F^2/2\big), \quad (2.1)$$

where $m = \prod_{k=1}^K m_k$ and $\mathbf{\Sigma}^{-1/2} := \{\mathbf{\Sigma}_1^{-1/2},\ldots,\mathbf{\Sigma}_K^{-1/2}\}$. When $K = 1$, this tensor normal distribution reduces to the vector normal distribution with zero mean and covariance $\mathbf{\Sigma}_1$. According to [47], it can be shown that $\mathcal{T} \sim \text{TN}(\mathbf{0};\mathbf{\Sigma}_1,\ldots,\mathbf{\Sigma}_K)$ if and only if $\text{vec}(\mathcal{T}) \sim \text{N}(\text{vec}(\mathbf{0});\mathbf{\Sigma}_K\otimes\cdots\otimes\mathbf{\Sigma}_1)$, where $\text{vec}(\mathbf{0}) \in \mathbb{R}^m$ and $\otimes$ is the matrix Kronecker product.

We consider the parameter estimation for the tensor normal model. Assume that we observe independently and identically distributed tensor samples $\mathcal{T}_1,\ldots,\mathcal{T}_n$ from $\text{TN}(\mathbf{0};\mathbf{\Sigma}_1^*,\ldots,\mathbf{\Sigma}_K^*)$. We aim to estimate the true covariance matrices $(\mathbf{\Sigma}_1^*,\ldots,\mathbf{\Sigma}_K^*)$ and their corresponding true precision matrices $(\mathbf{\Omega}_1^*,\ldots,\mathbf{\Omega}_K^*)$ where $\mathbf{\Omega}_k^* = \mathbf{\Sigma}_k^{*-1}$ $(k = 1,\ldots,K)$. To address the identifiability issue in the parameterization of the tensor normal distribution, we assume that $\|\mathbf{\Omega}_k^*\|_F = 1$ for $k = 1,\ldots,K$. This renormalization does not change the graph structure of the original precision matrix.

A standard approach to estimate $\mathbf{\Omega}_k^*$, $k = 1,\ldots,K$, is to use the maximum likelihood method via (2.1). Up to a constant, the negative log-likelihood function of the tensor normal distribution is

$$\ell(\mathbf{\Omega}_1,\ldots,\mathbf{\Omega}_K) := \frac{1}{2}\text{tr}[\mathbf{S}(\mathbf{\Omega}_K\otimes\cdots\otimes\mathbf{\Omega}_1)] - \frac{1}{2}\sum_{k=1}^K \frac{m}{m_k}\log|\mathbf{\Omega}_k|,$$

where $\mathbf{S} := \frac{1}{n}\sum_{i=1}^{n} \text{vec}(\mathcal{T}_i)\text{vec}(\mathcal{T}_i)^\top$. To encourage the sparsity of each precision matrix in the high-dimensional scenario, we propose a penalized log-likelihood estimator which minimizes $q_n(\mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_K) :=$

$$\frac{1}{m}\text{tr}[\mathbf{S}(\mathbf{\Omega}_K \otimes \cdots \otimes \mathbf{\Omega}_1)] - \sum_{k=1}^{K} \frac{1}{m_k} \log |\mathbf{\Omega}_k| + \sum_{k=1}^{K} P_{\lambda_k}(\mathbf{\Omega}_k), \tag{2.2}$$

where $P_{\lambda_k}(\cdot)$ is a penalty function indexed by the tuning parameter λ_k. In this paper, we focus on the lasso penalty [48] $P_{\lambda_k}(\mathbf{\Omega}_k) = \lambda_k \sum_{i \neq j} |[\mathbf{\Omega}_k]_{i,j}|$. The estimation procedure applies similarly to a broad family of penalty functions, for example, the SCAD penalty [49], the adaptive lasso penalty [50], the MCP penalty [51], and the truncated ℓ_1 penalty [52].

The penalized model from (2.2) is called the sparse tensor graphical model. It reduces to the m_1-dimensional sparse gaussian graphical model [53], [54], [55] when $K = 1$, and the sparse matrix graphical model [12], [13], [14], [15] when $K = 2$. Our framework generalizes them to fulfill the demand of capturing the graphical structure of the higher-order tensor-valued data.

2.3 Estimation

This section introduces the estimation procedure for the proposed sparse tensor graphical model. A computationally efficient algorithm is provided to alternatively estimate all precision matrices.

Recall that in (2.2), $q_n(\mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_K)$ is jointly non-convex with respect to $\mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_K$. Nevertheless, it turns out to be a bi-convex problem since $q_n(\mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_K)$ is convex in $\mathbf{\Omega}_k$ when the rest $K - 1$ precision matrices are fixed. The nice bi-convex property plays a critical role in our algorithm construction and its theoretical analysis in §3.

Based on the bi-convex property, we propose to solve this non-convex problem by alternatively updating one precision matrix with other matrices being fixed. Note that, for any $k = 1, \ldots, K$, minimizing (2.2) with respect to $\mathbf{\Omega}_k$ while fixing the rest $K - 1$ precision matrices is equivalent to minimizing

$$L(\mathbf{\Omega}_k) := \frac{1}{m_k}\text{tr}(\mathbf{S}_k\mathbf{\Omega}_k) - \frac{1}{m_k} \log |\mathbf{\Omega}_k| + \lambda_k \|\mathbf{\Omega}_k\|_{1,\text{off}}. \tag{2.3}$$

Here, $\mathbf{S}_k := \frac{m_k}{nm}\sum_{i=1}^{n} \mathbf{V}_i^k \mathbf{V}_i^{k\top}$, where $\mathbf{V}_i^k := \big[\mathcal{T}_i \times \{\mathbf{\Omega}_1^{1/2}, \ldots, \mathbf{\Omega}_{k-1}^{1/2}, \mathbb{1}_{m_k}, \mathbf{\Omega}_{k+1}^{1/2}, \ldots, \mathbf{\Omega}_K^{1/2}\}\big]_{(k)}$ with × the tensor product operation and $[\cdot]_{(k)}$ the mode-k matricization operation defined in §2.1. The result in (2.3) can be shown by noting that $\mathbf{V}_i^k = [\mathcal{T}_i]_{(k)}\big(\mathbf{\Omega}_K^{1/2} \otimes \cdots \otimes \mathbf{\Omega}_{k+1}^{1/2} \otimes \mathbf{\Omega}_{k-1}^{1/2} \otimes \cdots \otimes \mathbf{\Omega}_1^{1/2}\big)^\top$ according to the properties of mode-k matricization shown by [47]. Hereafter, we drop the superscript k of $\mathbf{V}_i^k$ if there is no confusion. Note that minimizing (2.3) corresponds to estimating vector-valued Gaussian graphical model and can be solved efficiently via the glasso algorithm [55].

The details of our Tensor lasso (Tlasso) algorithm are shown in Algorithm 1. It starts with a random initialization and then alternatively updates each precision matrix until it converges. In §3, we will illustrate that the statistical properties of the obtained estimator are insensitive to the choice of the initialization (see the discussion following Theorem 3.5). In our numerical experiments, for each $k = 1, \ldots, K$, we set the initialization of k-th precision matrix as $\mathbb{1}_{m_k}$, which leads to superior numerical performance.

Algorithm 1 Solve sparse tensor graphical model via Tensor lasso (Tlasso)

```
Input: Tensor samples $\mathcal{T}_1 \ldots, \mathcal{T}_n$, tuning parameters $\lambda_1, \ldots, \lambda_K$, max number of iterations $T$.
Initialize $\mathbf{\Omega}_1^{(0)}, \ldots, \mathbf{\Omega}_K^{(0)}$ randomly as symmetric and positive definite matrices and set $t = 0$.
Repeat:
$t = t + 1$.
For $k = 1, \ldots, K$:
    Given $\mathbf{\Omega}_1^{(t)}, \ldots, \mathbf{\Omega}_{k-1}^{(t)}, \mathbf{\Omega}_{k+1}^{(t-1)}, \ldots, \mathbf{\Omega}_K^{(t-1)}$, solve (2.3) for $\mathbf{\Omega}_k^{(t)}$ via glasso.
    Normalize $\mathbf{\Omega}_k^{(t)}$ such that $\|\mathbf{\Omega}_k^{(t)}\|_F = 1$.
End For
Until $t = T$.
Output: $\widehat{\mathbf{\Omega}}_k = \mathbf{\Omega}_k^{(T)}$ $(k = 1, \ldots, K)$.
```

3 Theory of Statistical Optimization

We first prove the estimation errors in terms of Frobenius norm, max norm, and spectral norm, and then provide the model selection consistency of the estimator output from the Tlasso algorithm. For compactness, we defer the proofs of theorems to supplementary material.

3.1 Estimation Error in Frobenius Norm

Based on the penalized log-likelihood in (2.2), we define the population log-likelihood function as $q(\mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_K) :=$

$$\frac{1}{m}\mathbb{E}\big\{\text{tr}\big[\text{vec}(\mathcal{T})\text{vec}(\mathcal{T})^\top(\mathbf{\Omega}_K \otimes \cdots \otimes \mathbf{\Omega}_1)\big]\big\} - \sum_{k=1}^{K} \frac{1}{m_k} \log |\mathbf{\Omega}_k|. \tag{3.1}$$

By minimizing $q(\mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_K)$ with respect to $\mathbf{\Omega}_k$, $k = 1, \ldots, K$, we obtain the population minimization function with the parameter $\mathbf{\Omega}_{[K]-k} := \{\mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_{k-1}, \mathbf{\Omega}_{k+1}, \ldots, \mathbf{\Omega}_K\}$, i.e.,

$$M_k(\mathbf{\Omega}_{[K]-k}) := \underset{\mathbf{\Omega}_k}{\text{argmin}}\, q(\mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_K). \tag{3.2}$$

Our first theorem shows an interesting result that the above population minimization function recovers the true parameter in only one iteration.

Theorem 3.1. *For any $k = 1, \ldots, K$, if $\mathbf{\Omega}_j$ $(j \neq k)$ satisfies $\text{tr}(\mathbf{\Sigma}_j^*\mathbf{\Omega}_j) \neq 0$, then the population minimization function in (3.2) satisfies $M_k(\mathbf{\Omega}_{[K]-k}) = m\big[m_k \prod_{j \neq k} \text{tr}(\mathbf{\Sigma}_j^*\mathbf{\Omega}_j)\big]^{-1}\mathbf{\Omega}_k^*$.*

Theorem 3.1 indicates that the population minimization function recovers the true precision matrix up to a constant in only one iteration. If $\mathbf{\Omega}_j = \mathbf{\Omega}_j^*, j \neq k$, then $M_k(\mathbf{\Omega}_{[K]-k}) = \mathbf{\Omega}_k^*$. Otherwise, after a normalization such that $\|M_k(\mathbf{\Omega}_{[K]-k})\|_F = 1$, the normalized population minimization function still fully recovers $\mathbf{\Omega}_k^*$. This observation suggests that setting $T = 1$ in Algorithm 1 is sufficient. Such a theoretical suggestion will be further supported by our numeric results.

In practice, when the population log-likelihood function (3.1) is unknown, we can approximate it by its sample version $q_n(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_K)$ defined in (2.2), which gives rise to the statistical estimation error. Similar as (3.2), we define the sample-based minimization function with parameter $\boldsymbol{\Omega}_{[K]-k} = \{\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_{k-1}, \boldsymbol{\Omega}_{k+1}, \ldots, \boldsymbol{\Omega}_K\}$ as

$$\widehat{M}_k(\boldsymbol{\Omega}_{[K]-k}) := \underset{\boldsymbol{\Omega}_k}{\operatorname{argmin}}\, q_n(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_K). \tag{3.3}$$

In order to derive the estimation error, it remains to quantify the statistical error induced from finite samples. The following two regularity conditions are assumed for this purpose.

Condition 3.2 (Bounded Eigenvalues). For any $k = 1, \ldots, K$, there is a constant $C_1 > 0$ such that,

$$0 < C_1 \le \lambda_{\min}(\boldsymbol{\Sigma}_k^*) \le \lambda_{\max}(\boldsymbol{\Sigma}_k^*) \le 1/C_1 < \infty,$$

where $\lambda_{\min}(\boldsymbol{\Sigma}_k^*)$ and $\lambda_{\max}(\boldsymbol{\Sigma}_k^*)$ refer to the minimal and maximal eigenvalue of $\boldsymbol{\Sigma}_k^*$, respectively.

Condition 3.2 has been commonly assumed in the precision matrix estimation literature in order to facilitate the proof of estimation consistency [56], [57], [58].

Condition 3.3 (Tuning). For any $k = 1, \ldots, K$ and some constant $C_2 > 0$, the tuning parameter λ_k satisfies $1/C_2\sqrt{\log m_k/(nmm_k)} \le \lambda_k \le C_2\sqrt{\log m_k/(nmm_k)}$.

Before characterizing the statistical error, we define a sparsity parameter for $\boldsymbol{\Omega}_k^*$, $k = 1, \ldots, K$. Let $\mathbb{S}_k := \{(i,j) : [\boldsymbol{\Omega}_k^*]_{i,j} \neq 0\}$. Denote the sparsity parameter $s_k := |\mathbb{S}_k| - m_k$, which is the number of nonzero entries in the off-diagonal component of $\boldsymbol{\Omega}_k^*$. For each $k = 1, \ldots, K$, we define $\mathbb{B}(\boldsymbol{\Omega}_k^*)$ as the set containing $\boldsymbol{\Omega}_k^*$ and its neighborhood for some sufficiently large radius $\alpha > 0$, i.e., $\mathbb{B}(\boldsymbol{\Omega}_k^*) :=$

$$\{\boldsymbol{\Omega} \in \mathbb{R}^{m_k \times m_k} : \boldsymbol{\Omega} = \boldsymbol{\Omega}^\top; \boldsymbol{\Omega} \succ 0; \|\boldsymbol{\Omega} - \boldsymbol{\Omega}_k^*\|_F \le \alpha\}. \tag{3.4}$$

Theorem 3.4. Suppose that Conditions 3.2 and 3.3 hold. For any $k = 1, \ldots, K$, the statistical error of the sample-based minimization function defined in (3.3) satisfies that, for any fixed $\boldsymbol{\Omega}_j \in \mathbb{B}(\boldsymbol{\Omega}_j^*)$ $(j \neq k)$,

$$\begin{aligned}\big\|\widehat{M}_k(\boldsymbol{\Omega}_{[K]-k}) &- M_k(\boldsymbol{\Omega}_{[K]-k})\big\|_F \\ &= O_P\left(\sqrt{\frac{m_k(m_k+s_k)\log m_k}{nm}}\right),\end{aligned} \tag{3.5}$$

where $M_k(\boldsymbol{\Omega}_{[K]-k})$ and $\widehat{M}_k(\boldsymbol{\Omega}_{[K]-k})$ are defined in (3.2) and (3.3), and $m = \prod_{k=1}^K m_k$.

Theorem 3.4 establishes the estimation error associated with $\widehat{M}_k(\boldsymbol{\Omega}_{[K]-k})$ for arbitrary $\boldsymbol{\Omega}_j \in \mathbb{B}(\boldsymbol{\Omega}_j^*)$ with $j \neq k$. In comparison, previous work on the existence of a local solution with desired statistical property only establishes theorems similar to Theorem 3.4 for $\boldsymbol{\Omega}_j = \boldsymbol{\Omega}_j^*$ with $j \neq k$. The extension to an arbitrary $\boldsymbol{\Omega}_j \in \mathbb{B}(\boldsymbol{\Omega}_j^*)$ involves nontrivial technical barriers. Specifically, we first establish the rate of convergence of the difference between a sample-based quadratic form and its expectation (Lemma S.1) via Talagrand's concentration inequality [59]. This result is also of independent interest. We then carefully characterize the rate of convergence of $\mathbf{S}_k$ defined in (2.3) (Lemma S.2). Finally, we develop (3.5) using the results for vector-valued graphical models developed by [60].

According to Theorem 3.1 and Theorem 3.4, we obtain the rate of convergence of the Tlasso estimator in terms of Frobenius norm, which is our main result.

Theorem 3.5. Assume that Conditions 3.2 and 3.3 hold. For any $k = 1, \ldots, K$, if the initialization satisfies $\boldsymbol{\Omega}_j^{(0)} \in \mathbb{B}(\boldsymbol{\Omega}_j^*)$ for any $j \neq k$, then the estimator $\widehat{\boldsymbol{\Omega}}_k$ from Algorithm 1 with $T = 1$ satisfies,

$$\|\widehat{\boldsymbol{\Omega}}_k - \boldsymbol{\Omega}_k^*\|_F = O_P\left(\sqrt{\frac{m_k(m_k+s_k)\log m_k}{nm}}\right), \tag{3.6}$$

where $m = \prod_{k=1}^K m_k$ and $\mathbb{B}(\boldsymbol{\Omega}_j^*)$ is defined in (3.4).

Theorem 3.5 suggests that as long as the initialization is within a constant distance to the truth, the Tlasso algorithm attains a consistent estimator after only one iteration. This consistency is insensitive to the initialization since the constant α in (3.4) can be arbitrarily large. In literature, [16] show that there exists a local minimizer of (2.2) whose convergence rate can achieve (3.6). However, it is unknown if their algorithm can find such a minimizer since there could be many other local minimizers.

A notable implication of Theorem 3.5 is that, when $K \ge 3$, the estimator from the Tlasso algorithm can achieve estimation consistency even if we only have access to one observation, i.e., $n = 1$, which is often the case in practice. To see it, suppose that $K = 3$ and $n = 1$. When the dimensions m_1, m_2, and m_3 are of the same order of magnitude and $s_k = O(m_k)$ for $k = 1, 2, 3$, all the three error rates corresponding to $k = 1, 2, 3$ in (3.6) converge to zero.

Theorem 3.5 implies that the estimation of the k-th precision matrix takes advantage of the information from the j-th way $(j \neq k)$ of the tensor data. Consider a simple case that $K = 2$ and one precision matrix $\boldsymbol{\Omega}_1^* = \mathbb{1}_{m_1}$ is known. In this scenario the rows of the matrix data are independent and hence the effective sample size for estimating $\boldsymbol{\Omega}_2^*$ is in fact nm_1. The optimality result for the vector-valued graphical model [11] implies that the optimal rate for estimating $\boldsymbol{\Omega}_2^*$ is $\sqrt{(m_2+s_2)\log m_2/(nm_1)}$, which is consistent with our result in (3.6). Therefore, the rate in (3.6) obtained by the Tlasso estimator is minimax-optimal since it is the best rate one can obtain even when $\boldsymbol{\Omega}_j^*$ $(j \neq k)$ were known. As far as we know, this phenomenon has not been discovered by any previous work in tensor graphical model.

Remark 3.6. For $K = 2$, our tensor graphical model reduces to matrix graphical model with Kronecker product covariance structure [12], [13], [14], [15]. In this case, the rate of convergence of $\widehat{\boldsymbol{\Omega}}_1$ in (3.6) reduces to $\sqrt{(m_1+s_1)\log m_1/(nm_2)}$, which is much faster than $\sqrt{m_2(m_1+s_1)(\log m_1 + \log m_2)/n}$ established by [13] and $\sqrt{(m_1+m_2)\log[\max(m_1,m_2,n)]/(nm_2)}$ established by [14]. In literature, [12] shows that there exists a local minimizer of the objective function whose estimation errors match ours. However, it is unknown if their estimator can achieve such convergence rate. On the other hand, our theorem confirms that our algorithm is able to find such estimator with an optimal rate of convergence.

3.2 Estimation Error in Max Norm and Spectral Norm

We next derive the estimation error in max norm and spectral norm. Trivially, these estimation errors are bounded by that in Frobenius norm shown in Theorem 3.5. To develop improved rates of convergence in max and spectral norms, we need to impose stronger conditions on true parameters.

We first introduce some important notations. Denote d_k as the maximum number of non-zeros in any row of the true precision matrices $\mathbf{\Omega}^*_k$, that is,

$$d_k := \max_{i\in\{1,\ldots,m_k\}} \big|\{j \in \{1,\ldots,m_k\} : [\mathbf{\Omega}^*_k]_{i,j} \neq 0\}\big|, \quad (3.7)$$

with $|\cdot|$ the set cardinality. For each covariance matrix $\mathbf{\Sigma}^*_k$, we define $\kappa_{\mathbf{\Sigma}^*_k} := |||\mathbf{\Sigma}^*_k|||_\infty$. Denote the Hessian matrix $\mathbf{\Gamma}^*_k := \mathbf{\Omega}^{*-1}_k \otimes \mathbf{\Omega}^{*-1}_k \in \mathbb{R}^{m_k^2 \times m_k^2}$, whose entry $[\mathbf{\Gamma}^*_k]_{(i,j),(s,t)}$ corresponds to the second order partial derivative of the objective function with respect to $[\mathbf{\Omega}_k]_{i,j}$ and $[\mathbf{\Omega}_k]_{s,t}$. We define its sub-matrix indexed by $\mathbb{S}_k$ as $[\mathbf{\Gamma}^*_k]_{\mathbb{S}_k,\mathbb{S}_k} = [\mathbf{\Omega}^{*-1}_k \otimes \mathbf{\Omega}^{*-1}_k]_{\mathbb{S}_k,\mathbb{S}_k}$, which is the $|\mathbb{S}_k| \times |\mathbb{S}_k|$ matrix with rows and columns of $\mathbf{\Gamma}^*_k$ indexed by $\mathbb{S}_k$ and $\mathbb{S}_k$, respectively. Moreover, we define $\kappa_{\mathbf{\Gamma}^*_k} := |||([\mathbf{\Gamma}^*_k]_{\mathbb{S}_k,\mathbb{S}_k})^{-1}|||_\infty$. In order to establish the rate of convergence in max norm, we need to impose an irrepresentability condition on the Hessian matrix.

Condition 3.7 (Irrepresentability). For each $k = 1,\ldots,K$, there exists some $\alpha_k \in (0,1]$ such that

$$\max_{e\in\mathbb{S}_k^c} \big\|[\mathbf{\Gamma}^*_k]_{e,\mathbb{S}_k}\big([\mathbf{\Gamma}^*_k]_{\mathbb{S}_k,\mathbb{S}_k}\big)^{-1}\big\|_1 \leq 1-\alpha_k.$$

Condition 3.7 controls the influence of the non-connected terms in $\mathbb{S}_k^c$ on the connected edges in $\mathbb{S}_k$. This condition has been widely applied for developing the theoretical properties of lasso-type estimator [43], [61], [62].

Condition 3.8 (Bounded Complexity). For each $k = 1,\ldots,K$, the parameters $\kappa_{\mathbf{\Sigma}^*_k}$ and $\kappa_{\mathbf{\Gamma}^*_k}$ are bounded and the parameter d_k in (3.7) satisfies $d_k = o\big(\sqrt{nm}/(m_k \log m_k)\big)$.

Theorem 3.9. Suppose Conditions 3.2, 3.3, 3.7 and 3.8 hold. Assume $s_k = O(m_k)$ for $k = 1,\ldots,K$ and assume $m_k' s$ are in the same order, i.e., $m_1 \asymp m_2 \asymp \cdots \asymp m_K$. For each k, if the initialization satisfies $\mathbf{\Omega}^{(0)}_j \in \mathbb{B}(\mathbf{\Omega}^*_j)$ for any $j \neq k$, then the estimator $\widehat{\mathbf{\Omega}}_k$ from Algorithm 1 with $T = 2$ satisfies,

$$\big\|\widehat{\mathbf{\Omega}}_k - \mathbf{\Omega}^*_k\big\|_\infty = O_P\left(\sqrt{\frac{m_k \log m_k}{nm}}\right). \quad (3.8)$$

In addition, the edge set of $\widehat{\mathbf{\Omega}}_k$ is a subset of the true edge set of $\mathbf{\Omega}^*_k$, that is, $\mathrm{supp}(\widehat{\mathbf{\Omega}}_k) \subseteq \mathrm{supp}(\mathbf{\Omega}^*_k)$.

Theorem 3.9 shows that the Tlasso estimator achieves the optimal rate of convergence in max norm [11]. Here we consider the estimator obtained after two iterations since we require a new concentration inequality (Lemma S.3) for the sample covariance matrix, which is built upon the estimator in Theorem 3.5.

Remark 3.10. Theorem 3.9 ensures that the estimated precision matrix correctly excludes all non-informative edges and includes all the true edges (i,j) with $|[\mathbf{\Omega}^*_k]_{i,j}| > C\sqrt{m_k \log m_k/(nm)}$ for some constant $C > 0$. Therefore, in order to achieve the variable selection consistency $\mathrm{sign}(\widehat{\mathbf{\Omega}}_k) = \mathrm{sign}(\mathbf{\Omega}^*_k)$, a sufficient condition is to assume that the minimal signal $\min_{(i,j)\in\mathrm{supp}(\mathbf{\Omega}^*_k)} |[\mathbf{\Omega}^*_k]_{i,j}| \geq C\sqrt{m_k \log m_k/(nm)}$ for each k. This confirms that the Tlasso estimator is able to correctly recover the graphical structure of each way of the high-dimensional tensor data.

A direct consequence from Theorem 3.9 is the estimation error in spectral norm.

Corollary 3.11. Suppose the conditions of Theorem 3.9 hold, for any $k = 1,\ldots,K$, we have

$$\big\|\widehat{\mathbf{\Omega}}_k - \mathbf{\Omega}^*_k\big\|_2 = O_P\left(d_k\sqrt{\frac{m_k \log m_k}{nm}}\right). \quad (3.9)$$

Remark 3.12. Now we compare our obtained rate of convergence in spectral norm for $K = 2$ with that established in the sparse matrix graphical model literature. In particular, [15] establishes the rate of $O_P\big(\sqrt{m_k(s_k \vee 1)\log(m_1 \vee m_2)/(nm_k)}\big)$ for $k = 1, 2$. Therefore, when $d_k^2 \leq (s_k \vee 1)$, which holds for example in the bounded degree graphs, our obtained rate is faster. However, our faster rate comes at the price of assuming the irrepresentability condition. Using recent advance in nonconvex regularization [63], we can actually eliminate the irrepresentability condition. We leave this to future work.

4 Tensor Inference

This section introduces a statistical inference procedure for sparse tensor graphical models. In particular, built on Tlasso algorithm, a consistent test statistic is constructed for hypothesis

$$H_{0k,ij} : [\mathbf{\Omega}^*_k]_{i,j} = 0 \qquad \text{v.s.} \qquad H_{1k,ij} : [\mathbf{\Omega}^*_k]_{i,j} \neq 0, \quad (4.1)$$

$\forall 1 \leq i < j \leq m_k$ and $k = 1,\ldots,K$. Also, to simultaneously test all off-diagonal entries, a multiple testing procedure is developed with false discovery rate (FDR) control.

4.1 Construction of Test Statistic

Without loss of generality, we focus on testing $\mathbf{\Omega}^*_1$. For a tensor $\mathcal{T} \in \mathbb{R}^{m_1\times\cdots\times m_K}$, denote $\mathcal{T}_{-i_1,i_2,\ldots,i_K} \in \mathbb{R}^{m_1-1}$ as the vector by removing the i_1-th entry of $\mathcal{T}_{:,i_2,\ldots,i_K}$. Given that $\mathcal{T}$ follows a tensor normal distribution (2.1), we have, $\forall i_1 \in m_1$, $\mathcal{T}_{i_1,i_2,\ldots,i_K} | \mathcal{T}_{-i_1,i_2,\ldots,i_K} \sim$

$$\mathrm{N}\left(-[\mathbf{\Omega}^*_1]^{-1}_{i_1,i_1}[\mathbf{\Omega}^*_1]_{i_1,-i_1}\mathcal{T}_{-i_1,i_2,\ldots,i_K} ; [\mathbf{\Omega}^*_1]^{-1}_{i_1,i_1}\prod_{k=2}^K [\mathbf{\Sigma}^*_k]_{i_k,i_k}\right). \quad (4.2)$$

Inspired by (4.2), our tensor graphical model can be reformulated into a linear regression. Specifically, for tensor sample $\mathcal{T}_l$, $l = 1,\ldots,n$, (4.2) implies that,

$$\mathcal{T}_{l;i_1,i_2,\ldots,i_K} = \mathcal{T}^\top_{l;-i_1,i_2,\ldots,i_K}\boldsymbol{\theta}_{i_1} + \xi_{l;i_1,i_2,\ldots,i_K}, \quad (4.3)$$

where regression parameter $\boldsymbol{\theta}_{i_1} = -[\mathbf{\Omega}^*_1]^{-1}_{i_1,i_1}[\mathbf{\Omega}^*_1]_{i_1,-i_1}$, and noise

$$\xi_{l;i_1,i_2,\ldots,i_K} \sim \mathrm{N}(0 ; [\mathbf{\Omega}^*_1]^{-1}_{i_1,i_1}\prod_{k=2}^K [\mathbf{\Sigma}^*_k]_{i_k,i_k}). \quad (4.4)$$

Let $\widehat{\mathbf{\Omega}}_1$ be an estimate of $\mathbf{\Omega}_1$ obtained from Tlasso algorithm. Naturally, a plug-in estimate of $\boldsymbol{\theta}_{i_1}$ follows, i.e.,

$\widehat{\boldsymbol{\theta}}_{i_1} = (\widehat{\theta}_{1,i_1}, \ldots, \widehat{\theta}_{m_1-1,i_1})^\top := -[\widehat{\boldsymbol{\Omega}}_1]_{i_1,i_1}^{-1}[\widehat{\boldsymbol{\Omega}}_1]_{i_1,-i_1}$. Denote a residual of (4.3) as $\widehat{\xi}_{l;i_1,i_2,\ldots,i_K} :=$

$$\mathcal{T}_{l;i_1,i_2,\ldots,i_K} - \bar{\mathcal{T}}_{i_1,i_2,\ldots,i_K} - (\mathcal{T}_{l;-i_1,i_2,\ldots,i_K} - \bar{\mathcal{T}}_{-i_1,i_2,\ldots,i_K})^\top \widehat{\boldsymbol{\theta}}_{i_1},$$

where $\bar{\mathcal{T}} = \sum_{l=1}^n \mathcal{T}_l/n$. Correspondingly, its sample covariance is, $\forall 1 \le i < j \le m_1$, $\widehat{\varrho}_{i,j} =$

$$\frac{m_1}{(n-1)m} \sum_{l=1}^{n} \sum_{i_2=1}^{m_2} \cdots \sum_{i_K=1}^{m_K} \widehat{\xi}_{l;i,i_2,\ldots,i_K} \widehat{\xi}_{l;j,i_2,\ldots,i_K}.$$

In light of (4.4), information of $[\boldsymbol{\Omega}_1^*]_{i,j}$ is encoded in $\widehat{\varrho}_{i,j}$. In this sense, a test statistic is proposed, i.e.,

$$\tau_{i,j} = \frac{\widehat{\varrho}_{i,j} + \mu_{i,j}}{\varpi}, \forall 1 \le i < j \le m_1. \tag{4.5}$$

Intuition of $\tau_{i,j}$ is extracting knowledge of $[\boldsymbol{\Omega}_1^*]_{i,j}$ from $\widehat{\varrho}_{i,j}$ via two-step correction. Notably, bias correction term $\mu_{i,j} := \widehat{\varrho}_{i,i}\widehat{\theta}_{i,j} + \widehat{\varrho}_{j,j}\widehat{\theta}_{j-1,i}$ reduces bias resulting from estimation error of $\widehat{\boldsymbol{\theta}}_i$ and $\widehat{\boldsymbol{\theta}}_j$. In addition, variance correction term

$$\varpi^2 := \frac{m \cdot \|\widehat{\mathbf{S}}_2\|_F^2 \cdots \|\widehat{\mathbf{S}}_K\|_F^2}{m_1 \cdot [\mathrm{tr}(\widehat{\mathbf{S}}_2)]^2 \cdots [\mathrm{tr}(\widehat{\mathbf{S}}_K)]^2}, \tag{4.6}$$

eliminates extra variation introduced by the rest $K-1$ modes (see (4.4)). Here $\widehat{\mathbf{S}}_k := \frac{m_k}{nm}\sum_{i=1}^n \widehat{\mathbf{V}}_i \widehat{\mathbf{V}}_i^\top$ is an estimate of $\boldsymbol{\Sigma}_k$, where $\widehat{\mathbf{V}}_i := [\mathcal{T}_i \times \{\widehat{\boldsymbol{\Omega}}_1^{1/2}, \ldots, \widehat{\boldsymbol{\Omega}}_{k-1}^{1/2}, \mathbb{1}_{m_k}, \widehat{\boldsymbol{\Omega}}_{k+1}^{1/2}, \ldots, \widehat{\boldsymbol{\Omega}}_K^{1/2}\}]_{(k)}$ with $\widehat{\boldsymbol{\Omega}}_k$ from Tlasso algorithm.

Theorem 4.1 establishes asymptotic normality of $\tau_{i,j}$. Symmetrically, such normality can be extended to the rest $K-1$ modes.

Theorem 4.1. Assume the same assumptions of Theorem 3.9, we have, under null (4.1),

$$\widetilde{\tau}_{i,j} := \sqrt{\frac{(n-1)m}{m_1 \widehat{\varrho}_{i,i}\widehat{\varrho}_{j,j}}}\tau_{i,j} \to \mathrm{N}(0;1)$$

in distribution, as $nm/m_1 \to \infty$.

Theorem 4.1 implies that, when $K \ge 2$, asymptotic normality holds even if we have a constant number of observations, which is often the case in practice. For example, let $n = 2$ and $m_1 \asymp m_2$, nm/m_1 still goes to infinity as m_1, m_2 diverges . This result reflects an interesting phenomenon specifically in tensor graphical models. Particularly, hypothesis testing for certain mode's precision matrix could take advantage of information from the rest modes in tensor data. As far as we know, this phenomenon has not been discovered by any previous work in tensor graphical models.

4.2 FDR Control Procedure

Though our test statistic enjoys consistency on single entry, simultaneously testing all off-diagonal entries is more of practical interest. Thus, in this subsection, a multiple testing procedure with false discovery rate (FDR) control is developed.

Given a thresholding level ς, denote $\varphi_\varsigma(\widetilde{\tau}_{i,j}) := \mathbb{1}\{|\widetilde{\tau}_{i,j}| \ge \varsigma\}$. Null is rejected if $\varphi_\varsigma(\widetilde{\tau}_{i,j}) = 1$. Correspondingly, false discovery proportion (FDP) and FDR are defined as

$$\mathrm{FDP} = \frac{|\{(i,j) \in \mathcal{H}_0 : \varphi_\varsigma(\widetilde{\tau}_{i,j}) = 1\}|}{|\{(i,j) : 1 \le i < j \le m_1, \varphi_\varsigma(\widetilde{\tau}_{i,j}) = 1\}| \vee 1},$$

and FDR $= \mathbb{E}(\mathrm{FDP})$. Here $\mathcal{H}_0 = \{(i,j) : [\boldsymbol{\Omega}_1^*]_{i,j} = 0, 1 \le i < j \le m_1\}$. A sufficient small ς is ideal that significantly enhances power, meanwhile controls FDP under a pre-specific level $\upsilon \in (0,1)$. In particular, the ideal thresholding value is

$$\varsigma_* := \inf\{\varsigma > 0 : \mathrm{FDP} \le \upsilon\}.$$

However, in practice, ς_* is not attainable due to unknown $\mathcal{H}_0$ in FDP. Therefore, we approximate ς_* by the following heuristics. Firstly, Theorem 4.1 implies that $P(\varphi_\varsigma(\widetilde{\tau}_{i,j}) = 1)$ is close to $2(1-\Phi(\varsigma))$ asymptotically. So the numerator of FDP is approximately $2(1-\Phi(\varsigma))|\mathcal{H}_0|$. Secondly, sparsity indicates that most entries are zero. Consequently, $|\mathcal{H}_0|$ is nearly $w := m_1(m_1-1)/2$. Under the above concerns, an approximation of ς_* is $\widehat{\varsigma} =$

$$\inf\left\{\varsigma > 0 : \frac{2(1-\Phi(\varsigma))w}{|\{(i,j) : i < j, \varphi_\varsigma(\widetilde{\tau}_{i,j}) = 1\}| \vee 1} \le \upsilon\right\}, \tag{4.7}$$

which is a trivial one-dimensional search problem.

Algorithm 2 Support recovery with FDR control for sparse tensor graphical models

```
Input: Tensor samples $\mathcal{T}_1 \ldots, \mathcal{T}_n$, $\{\widehat{\boldsymbol{\Omega}}_k\}_{k=1}^K$ from Algorithm 1, and a pre-specific level $\upsilon$.
Initialize: Support $\mathcal{S} = \emptyset$.
Compute test statistic $\widetilde{\tau}_{i,j}$, $\forall 1 \le i < j \le m_1$, defined in Theorem 4.1.
Compute thresholding level $\widehat{\varsigma}$ in (4.7).
If $\widetilde{\tau}_{i,j} > \widehat{\varsigma}$, $\forall 1 \le i < j \le m_1$, reject null hypothesis and set $\mathcal{S} = \mathcal{S} \cup \{(i,j),(j,i)\}$.
Output: $\mathcal{S} \cup \{(i,i) : 1 \le i \le m_1\}$.
```

Algorithm 2 describes our multiple testing procedure with FDR control for support recovery of $\boldsymbol{\Omega}_1^*$. Extension to the rest $K-1$ modes is symmetric. Clearly, FDR and FDP for $\boldsymbol{\Omega}_1^*$ from Algorithm 2 are

$$\mathrm{FDP}_1 = \frac{|\{(i,j) \in \mathcal{H}_0 : \varphi_{\widehat{\varsigma}}(\widetilde{\tau}_{i,j}) = 1\}|}{|\{(i,j) : 1 \le i < j \le m_1, \varphi_{\widehat{\varsigma}}(\widetilde{\tau}_{i,j}) = 1\}| \vee 1},$$

and $\mathrm{FDR}_1 = \mathbb{E}(\mathrm{FDP}_1)$. To depict their asymptotic behavior, two additional conditions are imposed related to size of true alternatives and sparsity.

Condition 4.2 (Alternative Size). Denote $\varpi_0^2 = m \cdot \|\boldsymbol{\Sigma}_2^*\|_F^2 \cdots \|\boldsymbol{\Sigma}_K^*\|_F^2 / (m_1 \cdot (\mathrm{tr}(\boldsymbol{\Sigma}_2^*) \cdots \mathrm{tr}(\boldsymbol{\Sigma}_K^*))^2)$. It holds that $\left|\{(i,j) : 1 \le i < j \le m_1, |[\boldsymbol{\Omega}_1^*]_{i,j}|/\sqrt{[\boldsymbol{\Omega}_1^*]_{i,i}[\boldsymbol{\Omega}_1^*]_{j,j}} \ge 4\sqrt{\varpi_0 m_1 \log m_1/((n-1)m)}\}\right| \ge \sqrt{\log\log m_1}$.

Condition 4.3 (Sparsity). For some $\rho < 1/2$ and $\gamma > 0$, there exists a positive constant C such that $\max_{1 \le i \le m_1} \left|\{j : 1 \le j \le m_1, j \ne i, |[\boldsymbol{\Omega}_1^*]_{i,j}| \ge (\log m_1)^{-2-\gamma}\}\right| \le C m_1^\rho$.

Notably, Condition 4.2 and 4.3 imply an interesting interplay between sparsity and number of true alternatives. In addition, Condition 4.2 is nearly necessary in the sense that FDR control for large-scale multiple testing fails if number of true alternatives is fixed [64]. Also, if $|\mathcal{H}_0| = o(w)$

(Condition 4.3 fails), most hypotheses would be rejected, and $\text{FDP}_1 \to 0$. Thus FDR control makes no sense anymore.

Theorem 4.4 characterizes asymptotic properties of FDP_1 and FDR_1. For simplicity, we denote $w_0 = |\mathcal{H}_0|$.

Theorem 4.4. Assume the same assumptions of Theorem 4.1, together with Condition 4.2 & 4.3. If $m_1 \le (nm/m_1)^r$ and $w_0 \ge cw$ for some positive constants r and c, we have

$$\text{FDP}_1 w/\upsilon w_0 \to 1, \text{ and } \quad \text{FDR}_1 w/\upsilon w_0 \to 1$$

in probability as $nm/m_1 \to \infty$.

Theorem 4.4 shows that our FDR control procedure is still valid even when sample size is constant and dimensionality diverges. Similar to Theorem 4.1, this phenomenon is specific to tensor graphical models.

Remark 4.5. Theorem 4.4 can be utilized to control FDR and FDP of testing Kronecker product $\mathbf{\Omega}_1^* \otimes \cdots \otimes \mathbf{\Omega}_K^*$. Consider a simple example with $K = 3$, denote f_1, f_2, f_3 as numbers of false discoveries of testing $\mathbf{\Omega}_1^*, \mathbf{\Omega}_2^*, \mathbf{\Omega}_3^*$ respectively, and d_1, d_2, d_3 as numbers of corresponding off-diagonal discoveries. FDP and FDR of testing $\mathbf{\Omega}_1^* \otimes \mathbf{\Omega}_2^* \otimes \mathbf{\Omega}_3^*$ are

$$\text{FDP}_c = \frac{\alpha_0(m_3 + d_3) + (\alpha - \alpha_0 + m_1 m_2) f_3}{[\prod_{k=1}^3 (d_k + m_k) - m_1 m_2 m_3] \vee 1},$$

and $\text{FDR}_c = \mathbb{E}(\text{FDP}_c)$, where $\alpha_0 = f_1(m_2 + d_2) + (d_1 - f_1 + m_1) f_2$ and $\alpha = (d_1 + m_1)(d_2 + m_2) - m_1 m_2$. In practice, values of f_k, $k \in \{1, 2, 3\}$, can be estimated by υd_k by Theorem 4.4, given that all precision matrices are sparse enough. Therefore, define

$$\tau = \frac{\alpha_0'(m_3 + d_3) + (\alpha - \alpha_0' + m_1 m_2)\upsilon d_3}{[\prod_{k=1}^3 (d_k + m_k) - m_1 m_2 m_3] \vee 1},$$

where $\alpha_0' = \upsilon d_1(m_2 + 2d_2) + (m_1 - \upsilon d_1)\upsilon d_2$. Similar arguments of Theorem 4.4 imply that $\text{FDP}_c/\tau \to 1$ and $\text{FDR}_c/\tau \to 1$.

5 Simulations

In this section, we demonstrate superior empirical performance of proposed estimation and inference procedures for sparse tensor graphical models. These procedures are implemented into R package **Tlasso**.

At first, we present numerical study of the Tlasso algorithm with iteration $T = 1$ and compare it with two alternative approaches. The first alternative method is graphical lasso (Glasso) approach [55] that applies to vectorized tensor data. This method ignores tensor structure of observed samples, and estimates Kronecker product of precision matrices $\mathbf{\Omega}_1^* \otimes \cdots \otimes \mathbf{\Omega}_K^*$ directly. The second alternative method is iterative penalized maximum likelihood method (P-MLE) proposed by [16]. This method iteratively updates each precision matrix by solving an individual graphical lasso problem while fixing all other precision matrices until a pre-specified termination condition $\sum_{k=1}^K \|\widehat{\mathbf{\Omega}}_k^{(t)} - \widehat{\mathbf{\Omega}}_k^{(t-1)}\|_F / K \le 0.001$ is met.

In the Tlasso algorithm, the tuning parameter for updating $\widehat{\mathbf{\Omega}}_k$ is set in the form of $C\sqrt{\log m_k/(nmm_k)}$ as assumed in Condition 3.3. Throughout all the simulations and real data analysis, we set $C = 20$. Sensitivity analysis in §S.4 of the online supplement shows that the performance of Tlasso is relatively robust to the value of C. For a fair comparison, the same tuning parameter is applied in P-MLE method for $k = 1, \ldots, K$. Individual graphical lasso problems in both Tlasso and P-MLE method are computed via *huge*. In the direct Glasso approach, its single tuning parameter is chosen by cross-validation automatically via *huge*.

In order to measure estimation accuracy of each method, three error criteria are selected. The first one is Frobenius estimation error of Kronecker product of precision matrices, i.e.,

$$\frac{1}{m}\left\|\widehat{\mathbf{\Omega}}_1 \otimes \cdots \otimes \widehat{\mathbf{\Omega}}_K - \mathbf{\Omega}_1^* \otimes \cdots \otimes \mathbf{\Omega}_K^*\right\|_F, \tag{5.1}$$

and the rest two are averaged estimation errors in Frobenius norm and max norm, i.e.,

$$\frac{1}{K}\sum_{k=1}^K \|\widehat{\mathbf{\Omega}}_k - \mathbf{\Omega}_k^*\|_F, \quad \frac{1}{K}\sum_{k=1}^K \|\widehat{\mathbf{\Omega}}_k - \mathbf{\Omega}_k^*\|_\infty. \tag{5.2}$$

Note that the last two criteria are only available to P-MLE method and Tlasso.

Two simulations are considered for a third order tensor, i.e., $K = 3$. In Simulation 1, we construct a triangle graph; in Simulation 2, a four nearest neighbor graph is adopted for each precision matrix. An illustration of generated graphs are shown in Figure 1. Detailed generation procedures for the two graphs are as follows.

Triangle: For each $k = 1, \ldots, K$, we construct covariance matrix $\mathbf{\Sigma}_k \in \mathbb{R}^{m_k \times m_k}$ such that its (i, j)-th entry is $[\mathbf{\Sigma}_k]_{i,j} = \exp(-|h_i - h_j|/2)$ with $h_1 < h_2 < \cdots < h_{m_k}$. The difference $h_i - h_{i-1}$, $i = 2, \ldots, m_k$, is generated i.i.d. from $\text{Unif}(0.5, 1)$. This generated covariance matrix mimics autoregressive process of order one, i.e., AR(1). We set $\mathbf{\Omega}_k^* = \mathbf{\Sigma}_k^{-1}$. Similar procedure has also been used by [60].

Nearest Neighbor: For each $k = 1, \ldots, K$, we construct precision matrix $\mathbf{\Omega}_k \in \mathbb{R}^{m_k \times m_k}$ directly from a four nearest-neighbor network. Firstly, m_k points are randomly picked from an unit square and all pairwise distances among them are computed. We then search for the four nearest-neighbors of each point and a pair of symmetric entries in $\mathbf{\Omega}_k$ corresponding to a pair of neighbors that has a randomly chosen value from $[-1, -0.5] \cup [0.5, 1]$. To ensure its positive definite property, the final precision matrix is designed as $\mathbf{\Omega}_k^* = \mathbf{\Omega}_k + (|\lambda_{\min}(\mathbf{\Omega}_k) + 0.2| \cdot \mathbb{1}_{m_k})$, where $\lambda_{\min(\cdot)}$ refers to the smallest eigenvalue. Similar procedure has also been studied by [65].

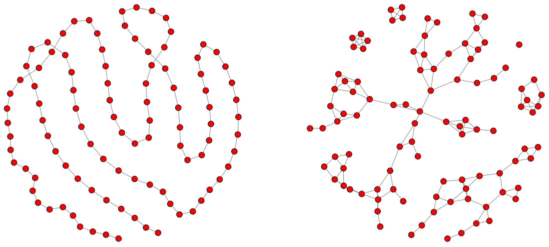


Fig. 1. An illustration of generated triangle graph (left) in Simulations 1 and four nearest neighbor graph (right) in Simulations 2. In this illustration, the dimension is 100.

In each simulation, we consider three scenarios as follows. Each scenario is repeated 100 times. Averaged computational time, and averaged criteria for estimation accuracy and variable selection consistency are computed.

- **Scenario s1:** sample size $n = 50$ and dimension $(m_1, m_2, m_3) = (10, 10, 10)$.
- **Scenario s2:** sample size $n = 80$ and dimension $(m_1, m_2, m_3) = (10, 10, 10)$.
- **Scenario s3:** sample size $n = 50$ and dimension $(m_1, m_2, m_3) = (10, 10, 20)$.

We first compare averaged computational time of all methods, see the first row of Figure 2. Clearly, Tlasso is dramatically faster than both competing methods. In particular, in Scenario s3, Tlasso takes about three seconds for each replicate. P-MLE takes about one minute while the direct Glasso method takes more than half an hour and is omitted in the plot. As we will show below, Tlasso algorithm is not only computationally efficient but also enjoys good estimation accuracy and support recovery performance.

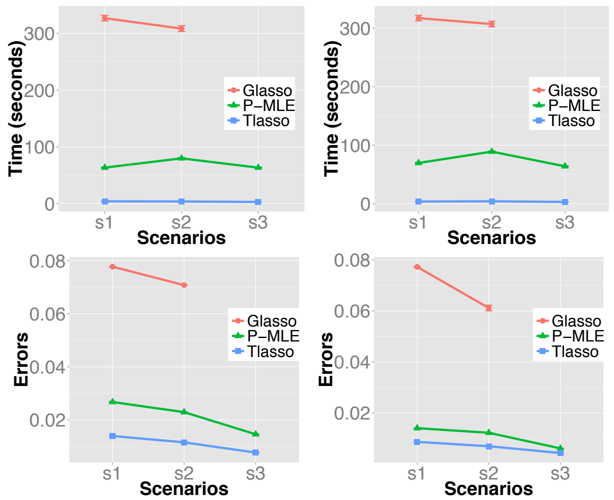


Fig. 2. The first row: averaged computational time of each method in Simulations $1\&2$, respectively. The second row: averaged estimation error of Kronecker product of precision matrices of each method in Simulations $1\&2$, respectively. Results for the direct Glasso method in Scenario s3 is omitted due to its extremely slow computation.

In the second row of Figure 2, we compute averaged estimation errors of Kronecker product of precision matrices. Clearly, with respect to tensor graphical structure, the direct Glasso method has significantly larger errors than Tlasso and P-MLE method. Tlasso outperforms P-MLE in Scenarios s1 and s2 and is comparable to P-MLE in Scenario s3. It is worth noting that, in Scenario s3, P-MLE is 20 times slower than Tlasso.

Next, we evaluate averaged estimation errors of precision matrices in Frobenius norm and max norm for Tlasso and P-MLE method. The direct Glasso method only estimate the whole Kronecker product, hence can not produce estimate for each precision matrix. Recall that, as we show in Theorem 3.5 and Theorem 3.9, estimation error for the k-th precision matrix is $O_P(\sqrt{m_k(m_k+s_k)\log m_k/(nm)})$ in Frobenius norm and $O_P(\sqrt{m_k \log m_k/(nm)})$ in max norm, where $m = m_1 m_2 m_3$ in this example. These theoretical findings are supported by numerical results in Figure 3. In particular, as sample size n increases from Scenario s1 to s2, estimation errors in both Frobenius norm and max norm expectedly decrease. From Scenario s1 to s3, one dimension m_3 increases from 10 to 20, and other dimensions m_1, m_2 keep the same, in which case averaged estimation error in max norm decreases, while error in Frobenius norm increases due to its additional $\sqrt{m_k + s_k}$ effect. Moreover, compared with P-MLE method, Tlasso demonstrates significant better performance in all three scenarios in terms of both Frobenius norm and max norm.

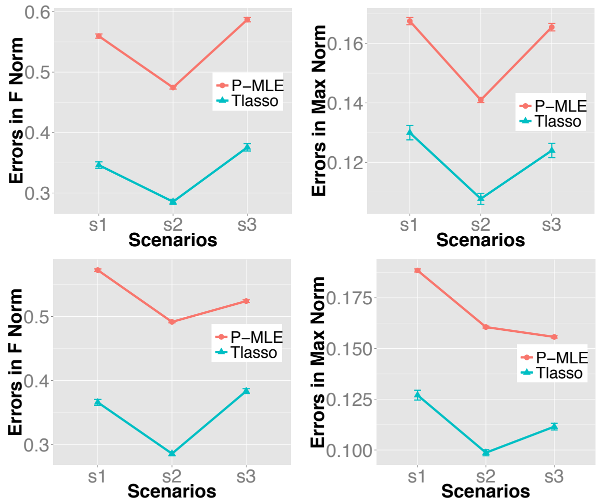


Fig. 3. Averaged estimation errors of precision matrices in Frobenius norm and max norm of each method in Simulations $1\&2$, respectively. The first row is for Simulation 1, and the second row is for Simulation 2.

From here, we turn to numerical study of the proposed inference procedure. Estimation of precision matrices in the inference procedure is conducted under the same setting as the former numerical study of Tlasso algorithm. Similarly, two simulations are considered, i.e., triangle graph and nearest neighbor graph. In both simulations, third-order tensors are constructed, adopting the same three scenarios as above: Scenario s1, s2, and s3. Each scenario repeats 100 times.

We first evaluate asymptotic normality of our test statistic $\widetilde{\tau}_{i,j}$. Figure 4 demonstrates QQ plots of test statistic for fixed zero entry $[\mathbf{\Omega}_1^*]_{6,1}$. Some other zero entries have been selected, and their simulation results are similar. So we only present results of $[\mathbf{\Omega}_1^*]_{6,1}$ in this section. As shown in Figure 4, our test statistic behaves very similar to standard normal even when sample size is small and dimensionality is high. It results from the fact that our inference method fully utilizes tensor structure.

Then we investigate the validity of our FDR control procedure. Table 1 contains FDP, its theoretical limit τ (see Remark 4.5), and power (all in $\%$) for Kronecker product of precision matrices under FDR control. Oracle procedure utilizes true covariance and precision matrices to compute test statistic. Each mode has the same pre-specific level $\upsilon = 5\%$ or 10%. As show in Table 1, powers are almost one and FDPs are small under poor conditions, i.e, small sample size or large dimensionality. It implies that our inference method has superior support recovery performance. Besides, empirical FDPs get closer to their theoretical limits if either of dimensionality and sample size is larger. This

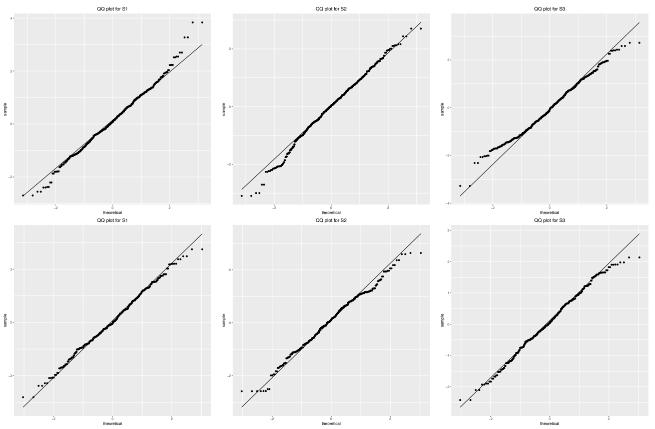


Fig. 4. QQ plots for fixed zero entry $[\mathbf{\Omega}_1^*]_{6,1}$. From left column to right column is scenario s1, s2 and s3. The first row is simulation 1, and the second is simulation 2.

phenomenon backs up the theoretical justification in Theorem 4.4. Thanks to fully utilizing tensor structure, difference between oracle FPR and our data-driven FDP decreases as either dimensionality or sample size grows.

TABLE 1
Empirical FDP, its theoretical limit τ, and power (all in %) of our inference procedure under FDP control for the Kronecker product of precision matrices in scenario s1, s2, and s3.

		Sim1			Sim2		
v		s1	s2	s3	s1	s2	s3
		Empirical FDP (τ)					
5	oracle	7.8	8.7	7.3	7.6	7.3	6.9
	data-driven	6.7 (9.9)	7.4 (9.9)	7.2 (9.9)	6.9 (11.1)	7 (11.1)	7.3 (11.1)
10	oracle	15.7	16.2	14.9	15.1	14.5	14.7
	data-driven	13.8 (19.3)	15.4 (19.3)	15.1 (19.4)	13.8 (21.4)	13.9 (21.4)	14.9 (21.4)
		Empirical Power					
5	oracle	100	100	100	99.9	100	99.9
	data-driven	100	100	100	99.8	100	99.8
10	oracle	100	100	100	100	100	100
	data-driven	100	100	100	99.9	100	99.9

In the end, we evaluate the true positive rate (TPR) and the true negative rate (TNR) of the Kronecker product of precision matrices for Glasso, P-MLE, and our FDP control procedure to compare their model selection performance. Specifically, let $a_{i,j}^*$ be the (i,j)-th entry of $\mathbf{\Omega}_1^* \otimes \cdots \otimes \mathbf{\Omega}_K^*$ and $\widehat{a}_{i,j}$ be the (i,j)-th entry of $\widehat{\mathbf{\Omega}}_1 \otimes \cdots \otimes \widehat{\mathbf{\Omega}}_K$, TPR and TNR of the Kronecker product are $\sum_{i,j} \mathbb{1}(\widehat{a}_{i,j} \neq 0, a_{i,j}^* \neq 0)/\sum_{i,j} \mathbb{1}(a_{i,j}^* \neq 0)$, and $\sum_{i,j} \mathbb{1}(\widehat{a}_{i,j} = 0, a_{i,j}^* = 0)/\sum_i \mathbb{1}(a_{i,j}^* = 0)$. Pre-specific FDP level is $v = 5\%$. Table 2 shows the model selection performance of all three methods. A good model selection procedure should produce large TPR and TNR. Our FDP control procedure has dominating TPR and TNR against the rest methods, i.e., almost all edges are identified and few non-connected edges are included.

TABLE 2
Model selection performance comparison among Glasso, P-MLE, and our FDP control procedure. Here TPR and TNR denote the true positive rate and true negative rate of the Kronecker product of precision matrices.

Scenarios	Glasso		P-MLE		Our FDP control	
	TPR	TNR	TPR	TNR	TPR	TNR
s1	0.343	0.930	1	0.893	1	0.935
Sim1 s2	0.333	0.931	1	0.894	1	0.932
s3	0.146	0.969	1	0.941	1	0.929
s1	0.152	0.917	1	0.854	0.999	0.926
Sim2 s2	0.119	0.938	1	0.851	1	0.926
s3	0.078	0.962	1	0.937	0.998	0.928

In short, the superior numerical performance and cheap computational cost in these simulations suggest that our method could be a competitive estimation and inferential tool for tensor graphical model in real-world applications.

6 Real Data Analysis

In this section, we apply our inference procedure on two real data sets. In particular, the first data set is from the Autism Brain Imaging Data Exchange (ABIDE), a study for autism spectrum disorder (ASD); the second set collects users' advertisement clicking behaviors from a major Internet company.

6.1 ABIDE

In this subsection, we analyze a real ASD neuroimaging dataset, i.e., ABIDE, to illustrate proposed inference procedure. As an increasingly prevalent neurodevelopmental disorder, symptoms of ASD are social difficulties, communication deficits, stereotyped behaviors and cognitive delays [66]. It is of scientific interest to understand how connectivity pattern of brain functional architecture differs between ASD subjects and normal controls. After preprocessing, ABIDE consists of the resting-state functional magnetic resonance imaging (fMRI) of 1071 subjects, of which 514 have ASD, and 557 are normal controls. fMRI image from each subject takes the form of a $30 \times 36 \times 30$-dimensional tensor of *fractional amplitude of low-frequency fluctuations* (fALFF), calculated at each brain voxels. In other words, ABIDE has 514+557 tensor images (each of dimension $30 \times 36 \times 30$) from ASD and controls, and these tensor images are 3D scans of human brain, whose entry values are fALFF of brain voxels at corresponding spatial locations. fALFF is a metric characterizing intensity of spontaneous brain activities, and thus quantifies functional architecture of the brain [67]. Therefore the support of precision matrix of fALFF fMRI images along each mode encodes the connectivity pattern of brain functional architecture. Dissimilarity in the supports between ASD and controls reveals potentially differential connectivity pattern. In this problem, vectorization methods, such as Glasso, will lose track of mode-specific structures, and thus can not be applied. Due to high dimensionality, false positive becomes a critical issue. However, P-MLE fails to guarantee FDP control as demonstrated in the simulation studies.

We apply the proposed inference procedure to recover the support of mode precision matrices of fALFF fMRI images of ASD group (514 image tensors) and normal control group (557 image tensors), respectively. Pre-specific FDP level is set as 0.01%. The rest setup is the same as in §5. Among the rejected entries of each group, we choose top 60 significant ones (smallest p-values) along each mode. All the selected entries show p-values less than 0.01%. Positions of differential entries between ASD and controls are recorded and mapped back to corresponding brain voxels. We further locate the voxels in the commonly used Anatomical Automatic Labeling (AAL) atlas [68], which consists of 116 brain regions of interest. Brain regions including the voxels, listed in Table 3, are suspected to have differential connectivity patterns between ASD and normal controls.

Our results in general match the established literature. For example, postcentral gyrus agrees with [69], which identifies postcentral gyrus as a key region where brain structure differs in autism. Also, [70] suggests that thalamus plays a role in motor abnormalities reported in autism studies. Moreover, temporal lobe demonstrates differential brain activity and brain volume in autism subjects [71].

TABLE 3
Brain regions of potentially differential connectivity pattern identified by our inference procedure.

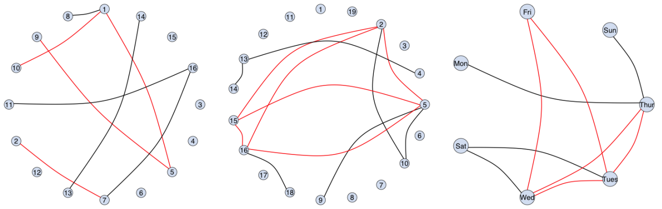
Hippocampus_L	ParaHippocampal_R	Hippocampus_R
Temporal_Sup_L	Amygdala_L	Temporal_Sup_R
Insula_L	Amygdala_R	Insula_R
Frontal_Mid_R	Thalamus_L	Thalamus_R
Pallidum_L	Putamen_L	Caudate_R
Precentral_L	Frontal_Inf_Oper_L	Frontal_Inf_Oper_R
Precentral_R	Postcentral_L	Postcentral_R
Temporal_Pole_Sup_R		

6.2 Advertisement Click Data

In this subsection, we apply the proposed inference method to an online advertising data set from a major Internet company. This dataset consists of click-through rates (CTR), i.e., the number of times a user has clicked on an advertisement from a certain device divided by the number of times the user has seen that advertisement from the device, for advertisements displayed on the company's webpages from May 19, 2016 to June 15, 2016. It tracks clicking behaviors of 814 users for 16 groups of advertisement from 19 publishers on each day of weeks, conditional on two devices, i.e., PC and mobile. Thus, two $16 \times 19 \times 7 \times 814$ tensors are formed by computing CTR corresponds to each (`advertisement`, `publisher`, `dayofweek`, `users`) quadruplet, conditional on PC and mobile respectively. However, more than 95% entries of either CTR tensor are missing. Hence, an alternating minimization tensor completion algorithm [72] is first conducted on the two tensors. Differential dependence structures within advertisements, publishers, and days of weeks between PC and mobile are of particular business interest. Therefore, we apply the proposed inference procedure to `advertisement`, `publisher`, and `dayofweek` modes of completed PC and mobile tensors respectively. Setup is the same as in §6.1. Among the rejected entries of each device, top $(30, 12, 10)$ significant ones in mode (`advertisement`, `publisher`, `dayofweek`) are selected. All the selected entries show p-values less than 0.01%. Pairs of entities, represented by the positions of differential entries between PC and mobile, are suspected to display dissimilar dependence when switching device.

Figure 5 demonstrates differential dependence patterns between PC and mobile in terms of advertisement, publishers, and days of weeks. Note that red lines indicate dependence only on PC, and black lines stand for those only on mobile. Due to confidential reason, description on specific entity of `advertisement` and `publisher` is not presented. We only provide general interpretations on the identified differential dependence patterns as follows. In `advertisement` mode, credit card ads and mortgage ads are linked on mobile. Such dependence is reasonable that people involved in mortgage would be more interested in credit card ads. On PC, uber share and solar energy are interchained. It can be interpreted in the sense that both uber share and solar energy are attractive for customers with energy-saving awareness. As for `publisher` mode, sport news publisher and weather news publisher are shown to be dependent on PC. This phenomenon can be accounted by the fact that sports and weather are the two most popular news choices when browsing websites. Also, magazine publishers (e.g., beauty magazines, tech magazines, and TV magazines) are connected on mobile. It is reasonable in the sense that people tend to read several casual magazines on mobiles for relaxing or during waiting. In `dayofweek` mode, strong dependence is demonstrated among weekdays, say from Tuesday to Friday, on PC. However, no clear pattern is showed on mobile. It can be explained that employees operate PC mostly at work on weekdays but use mobile every day.

Fig. 5. Analysis of the advertisement clicking data. Shown are differential dependence patterns between PC (red lines) and mobile (black lines) identified by our inference procedure. From left to right are advertisements, publishers, days of weeks.



7 Discussion

In this paper, we propose a novel sparse tensor graphical model to analyze graphical structure of high-dimensional tensor data. An efficient Tlasso algorithm is developed, which attains an estimator with minimax-optimal convergence rate in estimation. Tlasso algorithm not only is much faster than alternative approaches but also demonstrates superior estimation accuracy. In order to recover graph connectivity, we further develop an inference procedure with FDP control. Its asymptotic normality and validity of FDP control is rigorously justified. Numerical studies demonstrate its superior model selection performance. The above evidences motivate our methods more practically useful in comparison to other alternatives on real-life applications.

In Tlasso algorithm, graphical lasso penalty is applied for updating each precision matrix of tensor data. Lasso penalty is conceptually simple and computationally efficient. However, it is known to induce additional bias in estimation. In practice, other non-convex penalties, like SCAD [49], MCP [51], or Truncated ℓ_1 [52], are able to correct such bias. Optimization properties of non-convex penalized high-dimensional models have recently been studied by [73], which enables theoretical analysis of sparse tensor graphical model with non-convex penalties.

Acknowledgement

Han Liu is grateful for the support of NSF CAREER Award DMS1454377, NSF IIS1408910, NSF IIS1332109, NIH R01MH102339, NIH R01GM083084, and NIH R01HG06841. Guang Cheng's research is sponsored by NSF CAREER Award DMS-1151692, NSF DMS-1418042, DMS-1712907, Simons Fellowship in Mathematics, Office of Naval Research (ONR N00014-15-1-2331). Will Wei Sun was visiting Princeton and Guang Cheng was on sabbatical at Princeton while this work was carried out; Will Wei Sun and Guang Cheng would like to thank Princeton ORFE department for their hospitality and support.

References

[1] J. Jia and C.-K. Tang, "Tensor voting for image correction by global and local intensity alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 1, pp. 36–50, 2005.

[2] N. Zheng, Q. Li, S. Liao, and L. Zhang, "Flickr group recommendation based on tensor decomposition," in *International ACM SIGIR Conference*, 2010.

[3] A. Karatzoglou, X. Amatriain, L. Baltrunas, and N. Oliver, "Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering," in *ACM Recommender Systems*, 2010.

[4] S. Rendle and L. Schmidt-Thieme, "Pairwise interaction tensor factorization for personalized tag recommendation," in *International Conference on Web Search and Data Mining*, 2010.

[5] G. Allen, "Sparse higher-order principal components analysis," in *International Conference on Artificial Intelligence and Statistics*, 2012.

[6] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 208–220, 2013.

[7] Y. Wang, L. Yuan, J. Shi, A. Greve, J. Ye, A. W. Toga, A. L. Reiss, and P. M. Thompson, "Applying tensor-based morphometry to parametric surfaces can improve mri-based disease diagnosis," *Neuroimage*, vol. 74, pp. 209–230, 2013.

[8] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 208–220, 2013.

[9] P. Chu, Y. Pang, E. Cheng, Y. Zhu, Y. Zheng, and H. Ling, *Structure-Aware Rank-1 Tensor Approximation for Curvilinear Structure Tracking Using Learned Hierarchical Features*. Springer International Publishing, 2016, pp. 413–421.

[10] J. Zahn, S. Poosala, A. Owen, D. Ingram *et al.*, "AGEMAP: A gene expression database for aging in mice," *PLOS Genetics*, vol. 3, pp. 2326–2337, 2007.

[11] T. Cai, W. Liu, and H. Zhou, "Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation," *Annals of Statistics*, 2015.

[12] C. Leng and C. Tang, "Sparse matrix graphical models," *Journal of the American Statistical Association*, vol. 107, pp. 1187–1200, 2012.

[13] J. Yin and H. Li, "Model selection and estimation in the matrix normal graphical model," *Journal of Multivariate Analysis*, vol. 107, pp. 119–140, 2012.

[14] T. Tsiligkaridis, A. O. Hero, and S. Zhou, "On convergence of Kronecker graphical Lasso algorithms," *IEEE Transactions on Signal Processing*, vol. 61, pp. 1743–1755, 2013.

[15] S. Zhou, "Gemini: Graph estimation with matrix variate normal instances," *Annals of Statistics*, vol. 42, pp. 532–562, 2014.

[16] S. He, J. Yin, H. Li, and X. Wang, "Graphical model selection and estimation for high dimensional tensor data," *Journal of Multivariate Analysis*, vol. 128, pp. 165–185, 2014.

[17] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Symposium on Theory of Computing*, 2013, pp. 665–674.

[18] A. Agarwal, A. Anandkumar, P. Jain, and P. Netrapalli, "Learning sparsely used overcomplete dictionaries via alternating minimization," *arXiv:1310.7991*, 2013.

[19] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in *Advances in Neural Information Processing Systems*, 2013, pp. 2796–2804.

[20] X. Yi, C. Caramanis, and S. Sanghavi, "Alternating minimization for mixed linear regression," *arXiv:1310.3745*, 2013.

[21] S. Arora, R. Ge, and A. Moitra, "New algorithms for learning incoherent and overcomplete dictionaries," *arXiv:1308.6273*, 2013.

[22] M. Hardt, "Understanding alternating minimization for matrix completion," in *Symposium on Foundations of Computer Science*, 2014, pp. 651–660.

[23] M. Hardt, R. Meka, P. Raghavendra, and B. Weitz, "Computational limits for matrix completion," *arXiv:1402.2331*, 2014.

[24] M. Hardt and M. Wootters, "Fast matrix completion without the condition number," *arXiv:1407.4070*, 2014.

[25] S. Arora, A. Bhaskara, R. Ge, and T. Ma, "More algorithms for provable dictionary learning," *arXiv:1401.0579*, 2014.

[26] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere," *arXiv:1504.06785*, 2015.

[27] S. Arora, R. Ge, T. Ma, and A. Moitra, "Simple, efficient, and neural algorithms for sparse coding," *arXiv:1503.00778*, 2015.

[28] B. D. Haeffele and R. Vidal, "Global optimality in tensor factorization, deep learning, and beyond," *arXiv preprint arXiv:1506.07540*, 2015.

[29] Q. Sun, K. M. Tan, H. Liu, and T. Zhang, "Graphical nonconvex optimization for optimal estimation in gaussian graphical models," *arXiv preprint arXiv:1706.01158*, 2017.

[30] W. Chu and Z. Ghahramani, "Probabilistic models for incomplete multi-dimensional arrays," in *Artificial Intelligence and Statistics*, 2009, pp. 89–96.

[31] L. Xiong, X. Chen, T.-K. Huang, J. Schneider, and J. G. Carbonell, "Temporal collaborative filtering with bayesian probabilistic tensor factorization," in *Proceedings of the 2010 SIAM International Conference on Data Mining*. SIAM, 2010, pp. 211–222.

[32] P. Hoff, "Separable covariance arrays via the Tucker product, with applications to multivariate relational data," *Bayesian Analysis*, vol. 6, pp. 179–196, 2011.

[33] Z. Xu, F. Yan, and Y. Qi, "Infinite tucker decomposition: nonparametric bayesian models for multiway data analysis," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*. Omnipress, 2012, pp. 1675–1682.

[34] P. Rai, Y. Wang, S. Guo, G. Chen, D. Dunson, and L. Carin, "Scalable bayesian low-rank decomposition of incomplete multiway tensors," in *International Conference on Machine Learning*, 2014, pp. 1800–1808.

[35] P. D. Hoff *et al.*, "Equivariant and scale-free tucker decomposition models," *Bayesian Analysis*, vol. 11, no. 3, pp. 627–648, 2016.

[36] Q. Zhao, L. Zhang, and A. Cichocki, "Bayesian cp factorization of incomplete tensors with automatic rank determination," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1751–1763, 2015.

[37] C. Zhang and S. Zhang, "Confidence intervals for low dimensional parameters in high dimensional linear models," *Journal of the Royal Statistical Society, Series B*, vol. 76, pp. 217–242, 2014.

[38] S. van de Geer, P. Buhlmann, Y. Ritov, and R. Dezeure, "On asymptotically optimal confidence regions and tests for high-dimensional models," *Annals of Statistics*, vol. 42, pp. 1166–1202, 2014.

[39] A. Javanmard and A. Montanari, "De-biasing the lasso: Optimal sample size for gaussian designs," *arXiv preprint arXiv:1508.02757*, 2015.

[40] Y. Ning and H. Liu, "A general theory of hypothesis tests and confidence regions for sparse high dimensional models," *Annals of Statistics*, p. To Appear, 2016.

[41] X. Zhang and G. Cheng, "Simultaneous inference for high-dimensional linear models," *Journal of the American Statistical Association*, no. just-accepted, 2016.

[42] W. Liu, "Gaussian graphical model estimation with false discovery rate control," *The Annals of Statistics*, vol. 41, no. 6, pp. 2948–2978, 2013.

[43] J. Jankova and S. van de Geer, "Confidence intervals for high-dimensional inverse covariance estimation," *arXiv:1403.6752*, 2014.

[44] Z. Ren, T. Sun, C.-H. Zhang, and H. H. Zhou, "Asymptotic normality and optimalities in estimation of large gaussian graphical model," *Annals of Statistics*, p. To Appear, 2015.

[45] X. Chen and W. Liu, "Statistical inference for matrix-variate gaussian graphical models and false discovery rate control," *arXiv:1509.05453*, 2015.

[46] Y. Xia and L. Li, "Hypothesis testing of matrix graph model with application to brain connectivity analysis," *arXiv:1511.00718*, 2015.

[47] T. Kolda and B. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, pp. 455–500, 2009.

[48] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1996.

[49] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, pp. 1348—1360, 2001.

[50] H. Zou, "The adaptive Lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, pp. 1418—1429, 2006.

[51] C. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Annals of Statistics*, vol. 38, pp. 894—942, 2010.

[52] X. Shen, W. Pan, and Y. Zhu, "Likelihood-based selection and sharp parameter estimation," *Journal of the American Statistical Association*, vol. 107, pp. 223–232, 2012.

[53] M. Yuan and Y. Lin, "Model selection and estimation in the gaussian graphical model," *Biometrika*, vol. 94, pp. 19–35, 2007.

[54] O. Banerjee, L. Ghaoui, and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data," *Journal of Machine Learning Research*, vol. 9, pp. 485–516, 2008.

[55] J. Friedman, H. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical Lasso," *Biostatistics*, vol. 9, pp. 432–441, 2008.

[56] P. Bickel and E. Levina, "Covariance regularization by thresholding," *Annals of Statistics*, vol. 36, pp. 2577–2604, 2008.

[57] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu, "Sparse permutation invariant covariance estimation," *Electronic Journal of Statistics*, vol. 2, pp. 494–515, 2008.

[58] C. Lam and J. Fan, "Sparsistency and rates of convergence in large covariance matrix estimation," *Annals of Statistics*, vol. 37, pp. 4254–4278, 2009.

[59] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 2011.

[60] J. Fan, Y. Feng, and Y. Wu., "Network exploration via the adaptive Lasso and scad penalties," *Annals of Statistics*, vol. 3, pp. 521–541, 2009.

[61] P. Zhao and B. Yu, "On model selection consistency of Lasso," *Journal of Machine Learning Research*, vol. 7, pp. 2541–2567, 2006.

[62] P. Ravikumar, M. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," *Electronic Journal of Statistics*, vol. 5, pp. 935–980, 2011.

[63] P.-L. Loh and M. J. Wainwright, "Support recovery without incoherence: A case for nonconvex regularization," *arXiv:1412.5632*, 2014.

[64] W. Liu and Q.-M. Shao, "Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control," *Annals of Statistics*, vol. 42, no. 5, pp. 2003–2025, 2014.

[65] W. Lee and Y. Liu, "Joint estimation of multiple precision matrices with common structures," *Journal of Machine Learning Research*, p. To Appear, 2015.

[66] J. D. Rudie, J. Brown, D. Beck-Pancer, L. Hernandez, E. Dennis, P. Thompson, S. Bookheimer, and M. Dapretto, "Altered functional and structural brain network organization in autism," *NeuroImage: clinical*, vol. 2, pp. 79–94, 2013.

[67] R. Shi and J. Kang, "Thresholded multiscale gaussian processes with application to bayesian feature selection for massive neuroimaging data," *arXiv:1504.06074*, 2015.

[68] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, "Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain," *Neuroimage*, vol. 15, no. 1, pp. 273–289, 2002.

[69] K. L. Hyde, F. Samson, A. C. Evans, and L. Mottron, "Neuroanatomical differences in brain areas implicated in perceptual and other core features of autism revealed by cortical thickness analysis and voxel-based morphometry," *Human brain mapping*, vol. 31, no. 4, pp. 556–566, 2010.

[70] A. Nair, J. M. Treiber, D. K. Shukla, P. Shih, and R.-A. Müller, "Impaired thalamocortical connectivity in autism spectrum disorder: a study of functional and anatomical connectivity," *Brain*, vol. 136, no. 6, pp. 1942–1955, 2013.

[71] S. Ha, I.-J. Sohn, N. Kim, H. J. Sim, and K.-A. Cheon, "Characteristics of brains in autism spectrum disorder: Structure, function and connectivity across the lifespan," *Experimental neurobiology*, vol. 24, no. 4, pp. 273–284, 2015.

[72] P. Jain and S. Oh, "Provable tensor factorization with missing data," in *Advances in Neural Information Processing Systems*, 2014, pp. 1431–1439.

[73] Z. Wang, H. Liu, and T. Zhang, "Optimal computational and statistical rates of convergence for sparse nonconvex learning problems," *Annals of Statistics*, vol. 42, pp. 2164–2201, 2014.

[74] A. Gupta and D. Nagar, *Matrix variate distributions*. Chapman and Hall/CRC Press, 2000.

[75] A. Dawid, "Some matrix-variate distribution theory: Notational considerations and a bayesian application," *Biometrika*, vol. 68, pp. 265–274, 1981.

[76] J. Peng, P. Wang, N. Zhou, and J. Zhu, "Partial correlation estimation by joint sparse regression models," *Journal of the American Statistical Association*, 2012.

[77] T. Cai and W. Liu, "Adaptive thresholding for sparse covariance matrix estimation," *Journal of the American Statistical Association*, vol. 106, no. 494, pp. 672–684, 2011.

[78] S. Negahban and M. Wainwright, "Estimation of (near) low-rank matrices with noise and high-dimensional scaling," *Annals of Statistics*, vol. 39, pp. 1069–1097, 2011.

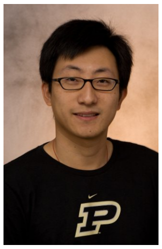
Xiang Lyu received his B.Economics degree from Renmin University, Beijing, China in 2016. He is currently working towards M.S. degree in Department of Statistics at Purdue University. His research interests include high-dimensional inference, tensor-valued data, graphical models, and non-convex optimization.

Will Wei Sun received BS degree in Statistics from Nankai University, China, in 2009, MS degree from University of Illinois at Chicago in 2011, and PhD degree from Purdue University in 2015. He then joined the advertising science team at Yahoo labs as a research scientist. He is currently an assistant professor in the Department of Management Science, University of Miami School of Business Administration, Florida. His research focuses on machine learning with applications in computational advertising, personalized recommendation system, and Neuroimaging analysis.

Zhaoran Wang will be joining Northwestern IEMS as an assistant professor in 2018. He works at the interface of machine learning, statistics, and optimization. He is the recipient of the AISTATS (Artificial Intelligence and Statistics Conference) notable paper award, ASA (American Statistical Association) best student paper in statistical learning and data mining, INFORMS (Institute for Operations Research and the Management Sciences) best student paper finalist in data mining, and the Microsoft fellowship.

Han Liu received a joint Ph.D. degree in Machine Learning and Statistics from the Carnegie Mellon University, Pittsburgh, PA, USA in 2011. He is currently the director of Reinforcement Learning Center at Tencent AI Lab. Beginning on September 1, 2018, he would be an Associate Professor of Electrical Engineering and Computer Science and Statistics at Northwestern University, Evanston, IL. He is also an adjunct Professor in the Department of Biostatistics and Department of Computer Science at Johns Hopkins University. From 2012-2017, he was an Assistant Professor of Statistical Machine Learning in the Department of Operations Research and Financial Engineering at Princeton University, Princeton, NJ. He built and is serving as the principal investigator of the Statistical Machine Learning (SMiLe) lab at Princeton University. His research interests include high dimensional semiparametric inference, statistical optimization, Big Data inferential analysis.

Jian Yang is a Senior Director of Advertising Sciences at Yahoo Research. He holds Ph.D. degree in Electrical and Computer Engineering from University of California, Davis. His research interests include optimization, forecasting and machine learning with applications in online advertising, pricing and revenue management, and supply chain management.



Guang Cheng received BA degree in Economics from Tsinghua University, China, in 2002, and PhD degree from University of Wisconsin–Madison in 2006. He then joined Dept of Statistics at Duke University as Visiting Assisitant Professor and Postdoc Fellow in SAMSI. He is currently Professor in Statisics at Purdue University, directing Big Data Theory research group, whose main goal is to develop computationally efficient inferential tools for big data with statistical guarantees.